



Aalto University
School of Science

CS-E4640 Big Data Platforms

Hands-on tutorial: Machine Learning Serving

Minh Tri Nguyen
Ph.D student of Aalto University
Researcher at AaltoSEA

Who I am? and what is this tutorial about?

■ Who I am?

- I am Minh Tri Nguyen
- MSc degree in Computer Science in 2019
- PhD student at Aalto University

■ What is this tutorial about

- An overview about Machine Learning (ML) Serving
- A quick demo of deploying a simple ML serving cluster.

Overview

▪ **Microservices architecture**

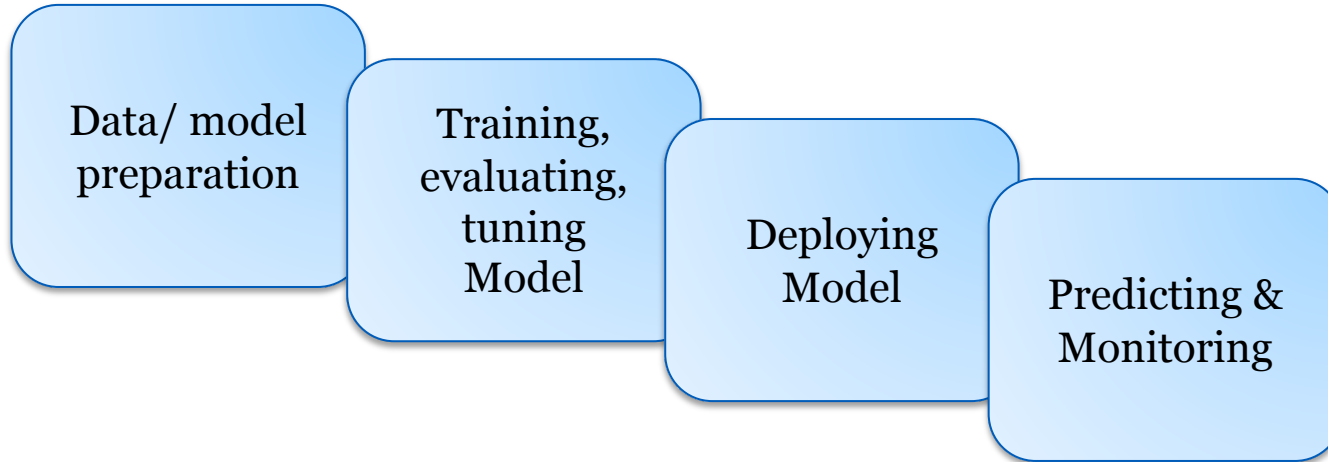
- Due to the high demand for diverse and flexible cloud service, microservices are emerging as a convenient way of deploying and managing software services. Instead of monolithic or linear logic block, the idea is to separate the whole service into microservices such as front-end, database, authentication & authorization, machine learning services, and so on. This approach allows modular programming at the microservices level, enabling code re-use and re-combine as needed.

▪ **Common commercial ML serving Platforms**

- Google AI Platform, Microsoft Azure, Prediction IO, ...

ML Serving

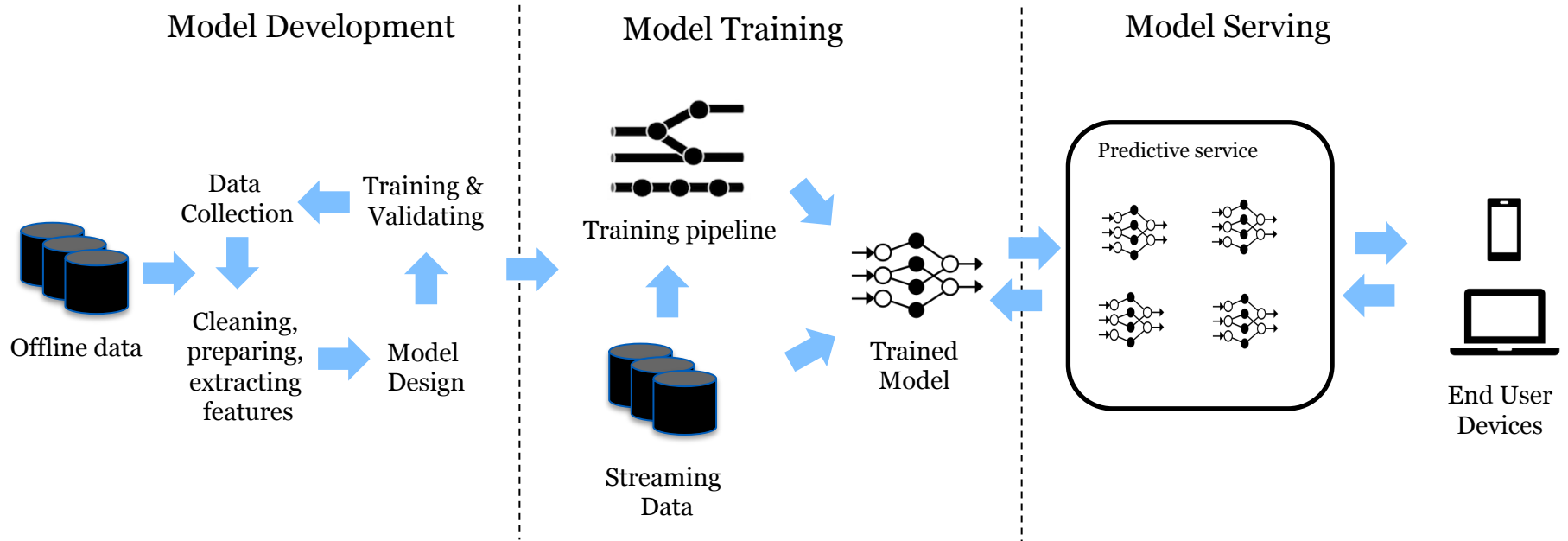
▪ ML Workflow



▪ Other stages

- Package model, update & version control...

ML Serving Lifecycle



Requirements for ML Serving

- **Performance**
 - Latency
 - Accuracy
 - ...
- **Scaling/Replicas**
- **Elasticity**
- **Cost**
- **Versioning models**
- Multiplex Models
- Batch processing

Approaches

- **Embed model in the web server**
 - Simple
 - End to end model control
 - Model load once, no isolation, no fine-grained replication
 - Pooling based process – memory issue with multi-model deployment
 - Hard to deploy complex pipeline

Approaches

- **Offload model to external service (cloud,...)**
 - Communication (API, ...)
 - Depend on cloud services (QoS, ...)
 - No infrastructure management
- **Private Cluster**
 - End to end model control
 - Privacy
- Separate service management
- Allow complex model deployment

A Quick Guide for ML Serving Cluster

- **Prerequisite**
 - Docker desktop
 - PredictionIO (docker, library)

Contact and References

- <https://version.aalto.fi/gitlab/bigdataplatforms/cs-e4640>
- <https://medium.com/retina-ai-health-inc/machine-learning-in-production-serving-up-multiple-ml-models-at-scale-the-tensorflow-serving-9607eeea30>
- <https://cloud.google.com/ai-platform/docs/technical-overview>
- https://youtu.be/rFZhwsSxZ_Q

Email: tri.m.nguyen@aalto.fi